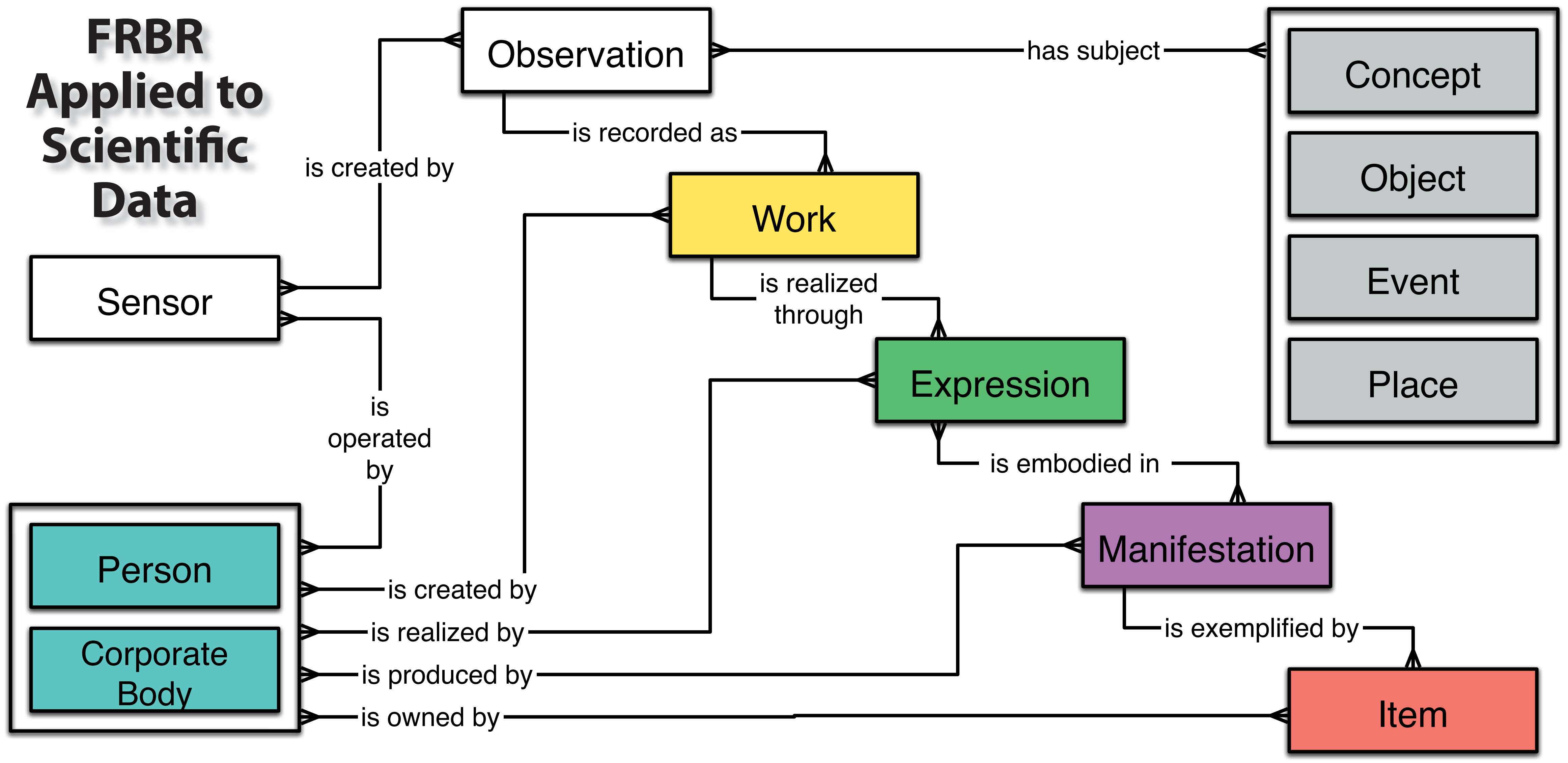
# Reconciling Heterogeneous Data Catalogs

Virtual Solar Observatory

Wyle Information Systems

http://www.virtualsolar.org/

J. A. Hourclé joseph.a.hourcle@nasa.gov



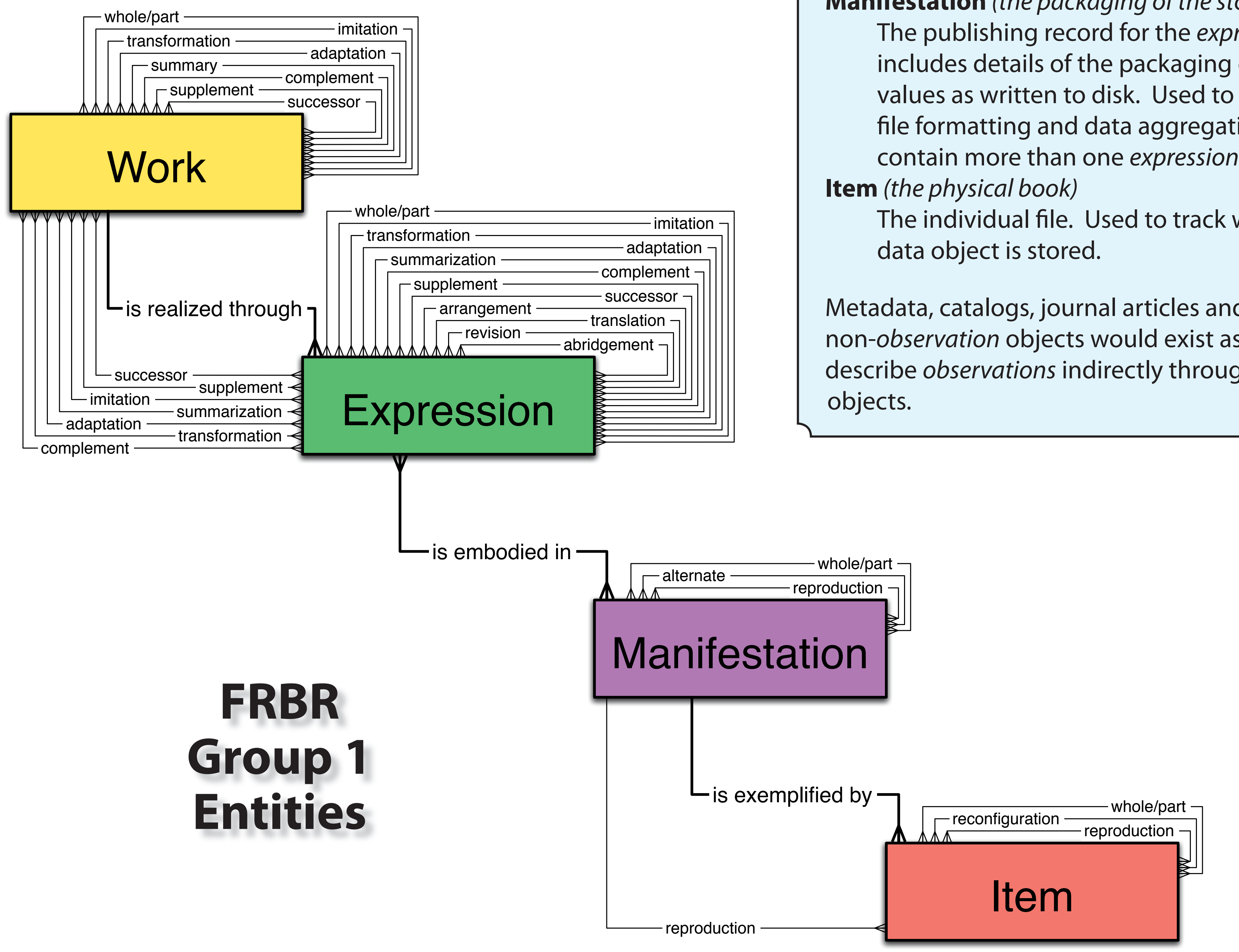
## Abstract

The term "data catalog" can be used to describe fundamentally different types of catalogs of an archive's holdings. We may track specific data files to be served by an archive or track a more abstract concept of data, such as the observation upon which the data file is based. If there is only one file for each observation, there may be no reason to distinguish between the two styles of cataloging.

If an archive serves multiple processed versions of each observation, or offers the data in more than one file format, the catalogs would be fundamentally different. Each style of cataloging serves a different purpose: the first allows an interested party to identify the exact calibration and packaging that they require, while the second may not. However, the second, a catalog of observations, allows scientists to identify an observation without being distracted by every permutation of processing and packaging.

The problems caused by differing catalog concepts becomes more apparent when building federated search systems. Unless the archives to be federated share the same concept of what their response records are, the merged results may vary from somewhat confusing to completely useless. This problem, however, is not restricted just to scientific data; library science has discussed the issue as the concept of "book" may be anything from the abstract creative work to a specific physical item. To assist in discussion, the library community developed the Functional Requirements for Bibliographic Records (IFLA 1998), a reference model that defines four entities that are commonly cataloged, as well as attributes and relationships for these and supporting entities.

We present an alignment of scientific data with FRBR and discuss how such a system framework provides improved usability of search systems, using examples encountered in development of the Virtual Solar Observatory.



FRBR Group 1 Entities

## FRBR as Applied to Scientific Data

We can reconcile the issue of observation vs. file catalogs by using a system that models multiple entities. The transformations typically applied to scientific data can be aligned with the entities in FRBR (Hourclé 2008):

**Observation** *(new object, not from FRBR)*
The data generated by a *sensor*. Used to track aspects of the pointing, location and observing mode of the *sensor*.

**Work** *(the abstract story within the book)*
The PI's interpretation (calibrated state) of the *observation*. Used to track flat fielding, conversion to physical units or other transformations to remove *sensor* effects.

**Expression** *(the words used to tell the story)*
The specific values used to express the *work*. Used to track other non-sensor specific transformations such as data compression, coordinate transformations, subsetting, binning and other types of data reduction.

**Manifestation** *(the packaging of the story)*
The publishing record for the *expression*; includes details of the packaging of the values as written to disk. Used to track file formatting and data aggregation; can contain more than one *expression*.

**Item** *(the physical book)*
The individual file. Used to track where the data object is stored.

Metadata, catalogs, journal articles and other non-*observation* objects would exist as *works* that describe *observations* indirectly through other objects.

## Disambiguation

The FRBR model identifies two additional task for catalog systems that come between the OAIS (CCSDS 2002) tasks of *finding* and *ordering*:

to **identify** an entity (ie, confirm that it corresponds to the entity sought, or to distinguish between multiple similar entities)

to **select** an entity that is appropriate to a user's needs

By using this model, we can more easily distinguish between the following types of "similar" files:

Two *observations* in sequence from the same *sensor*

Two *observations* with similar observing parameters taken from different *sensors*

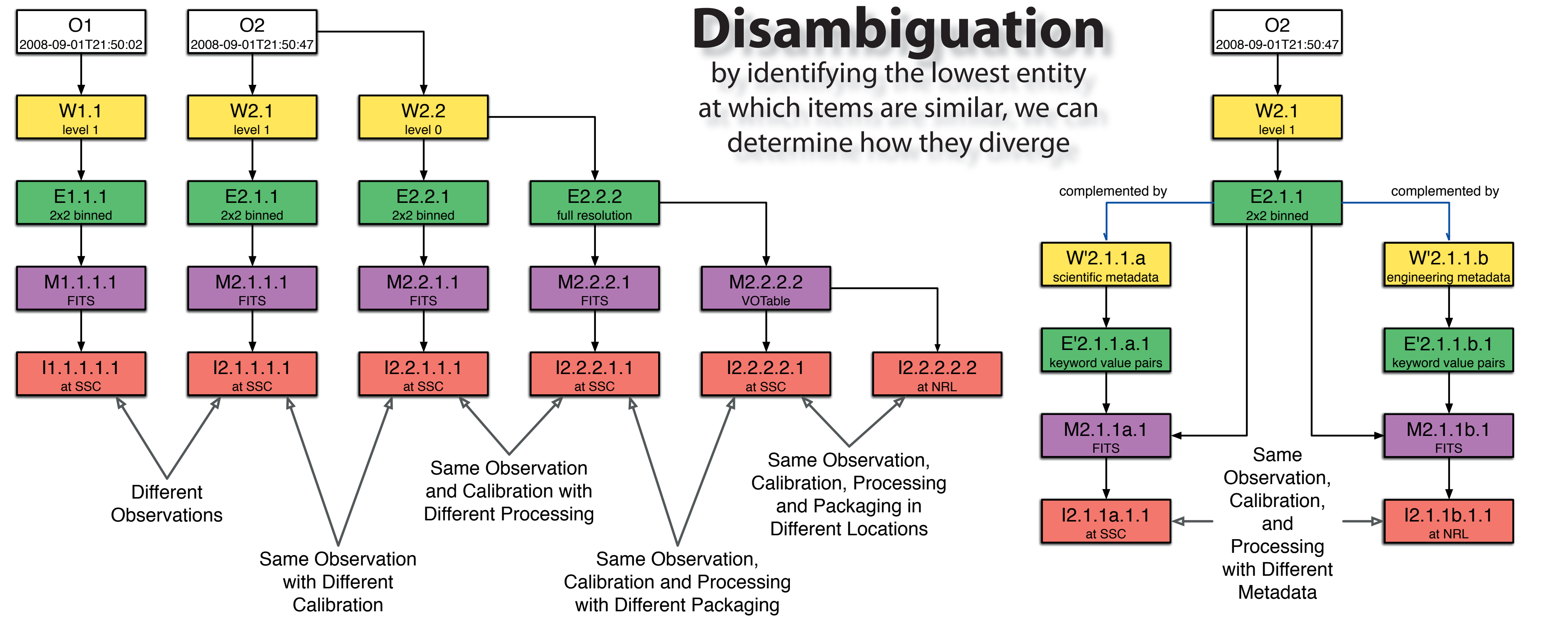
Two copies of the same *observation* with different calibration applied

Two copies of the same *observation* in different file formats

Two bytewise identical copies of the same file mirrored in different locations

Two copies of the same *observation*, processed in the same manner, using the same file format, but with different metadata attached.

We can then make decisions on how to handle duplicate records when presenting search results to the user—if files only differ by their location, and one is local to the user, there may be no reason to show two copies to the user. If we know what analysis tools are available, we may be able to determine that one file format is better for the user, and limit the amount of selections that they need to perform.



## Disambiguation

by identifying the lowest entity at which items are similar, we can determine how they diverge

## Practical Benefits

A multiple entity data model such as FRBR allows us to more specifically declare the relationships between files served by an archive. By assigning identifiers at the different entity levels, we can more easily track provenance and other relationships to allow researchers to ask questions to find the file that best serves their needs:

Is this *observation* available with level 1 calibration?

Is this *work* the most recent level 1 calibration available?

Is a browse image or plot available for this *expression*?

Is this *manifestation* available locally?

Is this *expression* available as FITS or NetCDF?

Are any forms of this *observation* available as FITS?

What is the next *manifestation* with similar observing mode, processing and packaging available from this *sensor*?

Where can I get the level 0 data for this *observation*?

If these identifiers are maintained across archives, we can associate higher level data objects with the PI's data objects on which they are based. We can also distinguish between files mirrored at a secondary archive versus other similar relationships previously mentioned.

It is hoped that this system can also be used to enable citations standards for scientific journals to identify the specific calibration and processing of the data used.

## Limitations

This model assumes that there are obvious boundaries between successive *observations* by a *sensor*. There are times when this is not true, or it is not practical to model. For instance, we would likely not track each individual value in a time series with a one second cadence, but might assign identifiers to each hour or day's worth of *observations*. We assume that the *observation* is at the smallest level of granularity practical for identification by an archive.

For systems that dynamically package their results, the *observation*, *work* and *expression* would remain fixed; there is no reason to track *manifestations* or *items*, as they do not exist until ordered by a user.

Some archives organize around the concept of a 'data series' or other data collections with a shared observing mode, calibration, processing and distribution format. Although this model could be used to better define the relationships between the series, it does not model the relationships between the individual files and the series as a whole. I have avoided this issue, as the nature of data series in solar physics results in a many-to-many relationship between *observations* and data series. The library community is looking into how to deal with a similar issue to reconcile catalogs of journals with the catalogs of articles they contain.

## References

Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. http://public.ccsds.org/publications/archive/650x0b1.pdf

Hourclé, J.A. (2007). FRBR Applied to Scientific Data. *Proc. ASIS&T 2008*, Columbus OH, USA. http://vso1.nascom.nasa.gov/vso/misc/jhourcle_ASIST_2008.pdf

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional Requirements for Bibliographic Records: final report.* München: K.G. Saur. http://www.ifla.org/VII/s13/frbr/frbr.pdf